International Society of the Learning Sciences

Sandboxing in Data Science: An Exploration of Youth Learning Using Open-Inquiry Approaches for Computing-based Data Mining

Alex Acquah, The University of Texas at El Paso, aacquah@miners.utep.edu
Amanda Barany, University of Pennsylvania, amanda.barany@gmail.com
Michael A. Johnson, University of North Texas, michaeljohnson12cbcl@my.unt.edu
Andi Scarola, The University of Texas at El Paso, adscarola@miners.utep.edu
Christopher Rivera, The University of Texas at El Paso, carivera7@miners.utep.edu
Justice T. Walker, The University of Texas at El Paso, jtwalker@utep.edu

Abstract: The integration of data science into high school computer science education is recognized for fostering productive learning with "big data." However, less attention has been placed on exercises where students generate datasets themselves rather than from sources curated by others, resulting in the need to explore learning designs that emphasize free inquiry. Few insights exist on designing and assessing learning outcomes in authentic practices, where learners lead inquiries and engagement processes. This research explores learning among eight youth participants in a two-week online computer science-based data mining workshop. Qualitative analysis of post-interviews and workshop artifacts reveals insights into youth learning during these activities. The findings highlight the significance of learning design strategies promoting active engagement in open-ended data science inquiries in computing education.

Introduction and background

Data science is increasingly recognized as an essential field across various disciplines, likened to the fourth industrial revolution for its role in innovation, occupational growth, and enhancing public literacy (Bureau of Labour Statistics, 2021). This recognition has spurred a movement to integrate data science into pre-college education, emphasizing the need for students to construct knowledge through data, beyond traditional data analytics (Lee & Wilkerson, 2018). However, current educational approaches often rely on pre-curated datasets, limiting the opportunity for students to engage in the authentic, multifaceted decision-making that characterizes real-world data science practice. This comprehensive approach, deeply rooted in the idea of 'situated cognition' (Greeno et al., 1996), suggests that learning should be embedded in authentic contexts, reflecting the complex nature of data science. This study explores the effectiveness of open inquiry in data science education through a workshop aimed at examining learning outcomes in authentic practice settings. To gain insights into these outcomes, we qualitatively examined their post-interview responses, addressing the following research questions: 1. What traditional or domain knowledge is fostered when employing an open-inquiry approach to computing-based data science? 2. What practice affordances are present in the context of computing-based data science open inquiry? Our findings suggest that embracing an open inquiry approach offers a robust framework for fostering effective computational data science.

Pre-college data science education aims to equip K-12 students with the foundational knowledge and skills for data analysis and interpretation, fostering data literacy (Lee & Delaney, 2022). Efforts to integrate data science and AI into education, like the bootstrap curriculum by Krishnamurthi et al. (2020), coding like a data miner—a culturally relevant curriculum (Barany et al., 2023; Walker et al., 2023), enhance mathematical and computing skills, while others advocate for a more humanistic approach to education (Lee et al., 2021; Wing, 2006). These initiatives address the urgent need for data proficiency in the modern world, emphasizing a holistic educational approach that incorporates ethical, cultural, and social awareness. However, disparities in educational access persist, highlighting the importance of developing curricula that are not only technically sound but also culturally and socially relevant (Jiang & Rosenberg, 2022).

Methods

This study explores data science learning outcomes through an online computer science workshop, engaging eight diverse participants aged 13–17 from the US and China. Using the "coding like a data miner curriculum," a sandbox curriculum focused on Twitter data mining with Python (Barany et al., 2023; Walker et al., 2023; Johnson et al., 2023). Over two weeks, participants engaged in the entire data science process, from mining to presenting findings. The analysis of post-interview transcripts from seven attendees, coded for data mining, heuristics, and computational thinking themes, employed inductive and deductive methods. Insights include varied hashtag use

for data collection, error diagnosis, and data cleaning in preprocessing, plus categorization and visualization adjustments in analysis. These outcomes illustrate how open inquiry can augment subject knowledge and equip high schoolers with hands-on data science skills.

Findings and implications

In an online workshop, participants utilized open inquiry for data science, significantly enhancing their computational thinking and mathematical domain knowledge. Through engaging with "big data," they demonstrated not only their proficiency in applying computational practices but also their deep understanding of mathematical principles within data science projects. For example, Janet's detailed account of her data science project process—from topic selection to employing codes for debugging and algorithm testing—illustrates a comprehensive grasp of computational methods and the critical, iterative nature of algorithm development. This scenario underscores the importance of debugging in the learning process, reflecting a nuanced understanding of code refinement. Several of the participants provided responses that evidenced their sense of debugging and often included the sense that reading or assessing code is an important step in this process (e.g., "[I would] probably try to fix the code first,then maybe like just find data manually, even though that's going to be like very hard. But the first option is just to try to fix the codes," (Vicento, 08/23/2023)). This indicates participants' adeptness at task decomposition, use of computational language, and debugging within data science contexts, illustrating their command over both computational practices and mathematical concepts. Notably, Joseph's handling of statistical outliers and Mira's diverse data collection exemplify a profound mathematical understanding pertinent to "big data." Such capabilities indicate that open inquiry not only promotes the integration of computational and domain knowledge but also encourages engagement with essential computer science practices.

Furthermore, the exploration of heuristic strategies reveals participants' adaptability, as seen in Mira's systematic data collection and Janet's thoughtful decision-making processes, which highlight the effectiveness of sandboxed learning environments. These environments facilitate a comprehensive approach to data science education, enhancing computational thinking and practical application skills.

Findings underscore the effectiveness of incorporating culturally relevant computing (CRC) into data science education, moving beyond traditional pedagogies to a more authentic, inclusive, and equitable framework. This approach significantly deepens students' understanding of computational and mathematical principles in practical scenarios. By promoting open inquiry and CRC, the findings suggest a potential transformation in pre-college data science education, underscoring the need for broader implementation and further research into peer interactions to enhance students' learning experiences.

Selected references

Barany, A., Reza, S., Johnson, M., Barrera, A., Badreddin, O., Fuentes, C., & Walker, J. T. (2023). Towards the Design of a Culturally Relevant Curriculum for Equitable, Data Mining-Based CS Education. In *Proceedings of the 17th International Conference of the Learning Sciences-ICLS* 2023, pp. 1498-1501.

Jiang, S., Lee, V. R., & Rosenberg, J. M. (2022). Data science education across the disciplines: Underexamined opportunities for K–12 innovation. *British Journal of Educational Technology, 53*(5), 1073-1079.

Johnson, M., Barany, A., Barrera, A., Acquah, A., and Walker, J. T. (2023). Lessons Learned from Online Codesign: Exploring Reflections of Connected Participatory Strategies for Computing-Based Data Science Curricular Design. In *Connected Learning Summit, Irvine, CA*, 2023.

Lee, V. R., Wilkerson, M. H., & Lanouette, K. (2021). A call for a humanistic stance toward K–12 data science education. Educational Researcher, 50(9), 664–672. https://doi.org/10.3102/0013189x211048810.

Walker, J. T., Barany, A., Acquah, A., Reza, S. M., Barrera, A., Del Rio Guzman, K., & Johnson, M. A. (2023). Coding like a data miner: A sandbox approach to computing-based data science for high school student learning. 2023 *IEEE Frontiers in Education Conference (FIE).*

Walker, J. T., Barany, A., Acquah, A., Reza, S. M., Guzman, K., Johnson, M., Badreldin, O., & Barrera, A. (2024). Sandbox Data Science: Culturally Relevant K-12 Computing. https://doi.org/10.1145/3631986

Acknowledgments

This work was supported in part by a grant from the National Science Foundation (#2137708) to Justice Walker, Amanda Barany, and Omar Badreddin. The views expressed are those of the authors and do not necessarily reflect those of the NSF or the University of Texas at El Paso.